ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Revisiting Statistical Applications in Soccer

STS Report

by Benoît Emonet

Benoit.Emonet@epfl.ch

Prof. A. C. Davison's Chair of Statistics
Department of Mathematics
Swiss Federal Institute of Technology
1015 Lausanne, Switzerland

Supervised by D. Kuonen and Prof. A. C. Davison

18th February 2000

Preface

The present report results from a project taking part of the 'Science, Technique and Society' cursus. These projects have to be done during the undergraduate studies in the Department of Mathematics at the Swiss Federal Institute of Technology (EPFL).

Our main goal was to review the statistical work related to soccer throughout articles published in statistical journals; see the references for an extensive list. To do so we merged them together in order to give an overall view of the different investigations.

Considering the quantity of articles written on the subject it was necessary to use a bibliography compilator. Therefore the second goal of our project was to get used with BibTeX, a LaTeX bibliography compiler.

Finally, as the importance to speak several languages is growing and as the mathematical literature is mainly written in English, the present report has been written in English as well.

Contents

1 Introduction

1.1 'Soccer' or 'Football'?

Football is called 'football' everywhere except in the United States, Australia and Canada where it is called 'soccer', and in Italy where it is called 'calcio', which is the name of a soccerlike game played in medieval Italy.

The name 'soccer' appeared when European people emigrated in the United States. Indeed, to distinguish this new sport with the football played there, it was called 'Association Football' and then shortened to 'Assoc. Football'. Some people just called it 'assoc.' or 'soc.'. Because it was common to add an 'er' to words at that time, 'soc.' became 'soccer'. The name stuck ever since.

In what follows, the word 'soccer' will be used in order to avoid any confusion with the football played for example in the United States.

1.2 History of soccer

Very early in history people began to play soccerlike games. There are records of such a game, called *tsu chu*, having been played in China more than 3000 years ago. Balls were made of animal skin and were kicked through a gap in a net stretched between poles of thirty feet high. It was played as a part of the emperor's birthday celebration and also to train soldiers during the Ts'in Dynasty (255–206 BC). The Munich Ethnological Museum in Germany has a Chinese text from approximatively 50 BC that mentions games played between Japan and China. But it is known for sure that a game was played in 611 in the ancient Japanese's capital of Kyoto.

In ancient Greece a game called *harpaston* was played but there are no additional description of it, except illustrations like the one in the left panel of Figure 1. The Romans played a game known as *harpastum* (right panel of Figure 1) which was probably brought by the Greeks (the names are very similar) and was the origin of modern soccer. They used a ball similar to the softball one, hence smaller and harder than the soccer one. It may have been very violent because during the Olympic games in ancient Rome twenty-seven men on each side competed so vigorously that two-thirds of them were hospitalised after a fifty minutes game. But as it exercised every part of the body the Romans liked it.

Alaska and Canada played a game on ice called *aqsaqtuk*. The balls were stuffed with grass, caribou hair and moss. A legend tells of two villages playing against each other with goals ten miles apart, but it is not known when it took place.

Finally the game arrived in England were it acquired a bad enough reputation among British royalty so that the government passed laws against soccer. For example, King Edward of England (1307–1327) proclaimed *'For as much as there is a great noise in the city caused hustling over large balls, from which many evils may arise, which God forbid, we command and forbid on behalf of the King, on pain of*

Figure 1: On the left a Greek is juggling with a ball and on the right some Romans are playing *harpastum*.

imprisonment, such game to be used in the city future.' Beside the fact that laws failed to stop the sport, the game became so popular by 1800 that, in certain annual contests in northern and middle England, large groups roamed and ragged through towns and villages.

In 1863 the 'Football Association' (FA) was created and the FA finally established uniform rules. After that everything went faster. In 1872 the first international game was played between England and Scotland (without counting the match between China and Japan in 50 BC). And it was in England where soccer professionalism was legalised in 1885. From there the game spread throughout the rest of the world.

By 1904 an international governing body was established to control the sport — the 'Fédération Internationale de Football Association' (FIFA). Today the FIFA, headquartered in Zürich (Switzerland), has more than 140 member nations and oversees the activities of about 39 million players worldwide.

The European part of the FIFA — the 'Union des Associations Européennes de Football' (UEFA) — was founded in Basel (Switzerland) in 1954 to organise the European competitions. Today the UEFA is composed of some 27 standing committees and oversees around 1'000 matches each season.

1.3 Soccer and Statistics

Considering the importance of this sport in the world, statisticians from countries like United Kingdom, Germany, United States, Canada and others started to create statistical models in order to predict the outcome of soccer games or to determine the best playing strategies. The goal was, and is always, to perform better predictions than the bookmakers.

In our knowledge, the first statistical considerations go back to Moroney (1956).

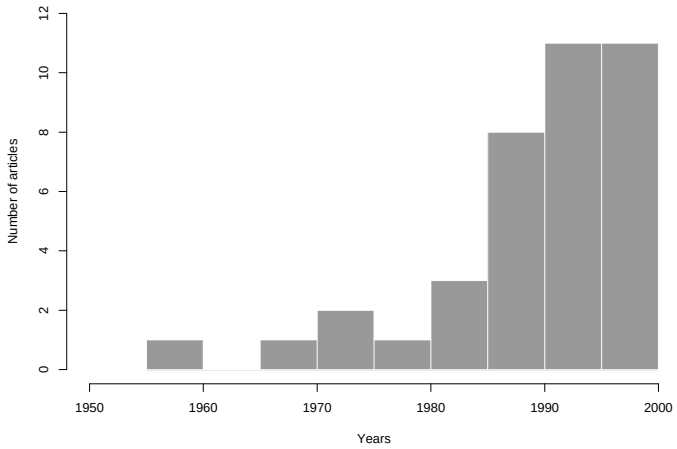

Figure 2: A histogram of the number of articles about soccer published in statistical journals. The data are grouped by intervals of 5 years.

But since then the number of articles published in specialised journals has increased significantly (see Figure 2). Indeed, while between 1950 and 1970 only 2 articles appeared, there were 11 between 1990 and 1995. This is certainly due to the growing popularity of soccer and perhaps also to the birth of powerful computers, which permit to calculate very complicated models; see also Section 3.

Hence the purpose of the present report is to provide an overall picture of statistical considerations related to soccer in order to show to someone wanting to search for such methods what has already been done and to create a database from which anybody could find references about articles he is interested in. To help this the interested reader can find its online version at:

`http://statwww.epfl.ch/projects/emonet`

1.4 Outline

This report is organised as follows. Section 2 presents the first models and developments as well as general considerations like the causes of home ground advantage or the effectiveness of different playing strategies. Section 3 reviews models made specifically for Round Robin tournaments, whereas Section 4 considers Knock Out tournaments. Knock Out tournaments with preliminary Round Robin stage, like the FIFA World Cup, are analysed in Section 5.

After some concluding remarks, the interested reader will find a list of web resources in the Appendix, together with an extensive list of references.

2 General considerations

2.1 Introduction

This section presents the methods and developments found until the early 1980's in the context of predicting the outcome of a soccer match. It contains also some considerations like the role of chance in a match, the causes of home ground advantage or the best playing strategies.

2.2 The first methods

The first statistician who presented a statistical model to predict the outcome of a soccer match was Moroney (1956). He found that the probability of winning a game is close to a Poisson and even closer to a negative binomial distribution.

But the most referred article, therefore certainly the most important one initially, is Reep and Benjamin (1968). The authors did not study directly the probability of winning a soccer match, but the probability $P(r)$ to do r-pass movements during a match. A r-pass move is a succession of r passes (between players from the same team) interrupted by one or more players from the opposing team. After preliminary studies on some data concerning these movements, they found that the distribution of $P(r)$ fits closely to a negative binomial distribution.

Some years later Reep et al. (1971) made a correction of their first model. Indeed they found that goal shots not arising from passing moves, but from penalties or interception, were erroneously included in the count of 0-pass moves. These are 0-pass moves in a certain sense, but are not really passes and they are not terminated by an interception of the opposing team. Hence they do not belong to the definition of 0-pass moves. After having excluded these shots, the fit to the negative binomial distribution was improved significantly. They conclude their paper by fitting the negative binomial distribution to the number of goals scored by a team during a match and agreed with the good fit obtained by Moroney (1956).

2.3 The role of chance

An important question which the statisticians tried to answer at the beginning was: is soccer a game of pure chance?

Using their study of r-pass moves, Reep and Benjamin (1968) answered that it seems that *'chance does dominate the game'*. But in their second work, Reep et al. (1971), they moderate their statement telling that skill supplants chance when the player is in the shooting area, where he has to choose between passing or shooting.

It is evident for anybody watching soccer games that for a single match, chance takes an important part, like the ball hitting the crossbar, a player slipping when he is in position to score or a favourable rebound. But considering a whole season, chance is certainly less important.

Hence instead of considering the role of chance in a single game, let it estimate throughout an entire season. Hill (1974) compared the tables made by expert forecasts before the beginning of the 1971–72 season to the tables of the final season's results: he did it for six different divisions (the first four English Football League divisions and the first two Scottish League divisions) and found six positive correlations which he interpreted as skills in both forecasts and players. He proved statistically that the role of skill throughout an entire season supplants the role of chance. Hence the idea to find a statistical model which could predict the outcome of soccer matches (in long runs rather then in short ones) has a sense, which it would not have, if chance had been more important because there is no way to modelise chance.

2.4 The developments

The models created by Moroney (1956), Reep and Benjamin (1968) or Reep et al. (1971) were not sufficiently developed to predict the outcome of a match between specific teams, accounting for the different quality of the teams involved. Maher (1982) was probably the first to propose such a model. His assumption was that the number of goals scored in a game by the home and away teams are independent Poisson variables. The means of these variables depend on the attack and defence qualities of each team. Explicitly, denote by i and j the home and away team. Let $X_{i,j}$ and $Y_{i,j}$ be the number of goals they scored in the match i against j. Then

$$\begin{aligned} X_{i,j} &\sim \text{Poisson}(\alpha_i \beta_j \gamma) \text{ and} \\ Y_{i,j} &\sim \text{Poisson}(\alpha_j \beta_i), \end{aligned}$$

where $X_{i,j}$ and $Y_{i,j}$ are independent variables, α_i and β_i measure their attack and defence qualities, $\forall\, i \neq j$. The home ground advantage of team i is measured by γ.

This model is more realistic than the models who tried to fit a particular distribution at a game outcome; see Moroney (1956). As each team has a personal attack and defence rate, it is possible to quantify the differences between stronger and weaker teams. Maher (1982) also gave a home advantage γ in order to modelise the advantage of playing on home ground.

The existence of a home advantage has been the goal of several articles reviewed in Courneya and Carron (1992). They observed the existence of such an advantage but were not able to determine the causes of home ground advantage.

2.5 What causes home ground advantage?

Pollard (1986) tried to answer this question quantifying home ground advantage as the number of matches won by home teams as a percentage of all games played (in a competition where every teams play an equal number of home and away matches). He found that there is a difference of home ground advantage between the different

divisions of a soccer league: the value of the home ground advantage decreased with the importance of the league.

This approach is quite acceptable if each team has approximately the same skills, otherwise, the difference of skills would distort the results. Indeed a strong team will frequently beat a weak one independent of home ground advantage. Therefore the model proposed by Pollard (1986) has to be improved in order to overcome such problems. This enables the need of a model to evaluate the teams ability in order to negate the effect caused by the different skills of each team. Clark and Norman (1995) and Kuk (1995) worked in this direction using a model like the one developed by Stefani (1980) for American football. It is a model based on a least squares system as described in what follows: Let w_{ij} be the goal margin in a match between team i and team j, with team i playing at home. Then

$$w_{ij} = u_i - u_j + h_i + \epsilon_{ij},$$

where u_i is a measure of team i's ability, h_i is a measure of team i's home ground advantage and ϵ_{ij} is a zero-mean random error, $\forall\, i \neq j$.

Using a least squares approach to simulate u_i and h_i, Clark and Norman (1995) obtained several new results concerning the causes of home ground advantage based on 94 teams from the English League during the years 1981–90.

- They noticed that the importance of the advantage to play at home depends on the years. Indeed, in 1982, 1983 and 1985 the importance of home advantage was 10% higher than in average and in 1981, 1987 and 1989 it was 10% lower.

- Considering h_{ij}, the paired home ground advantage of i when it plays versus j without considering the matches versus the other teams, they concluded that the geographical distance between the clubs may also be a cause of advantage (see Figure 3).

- Instead of w_{ij} being a goal margin let w_{ij} be 1, 0 or -1, depending on whether team i wins, draws or looses. They found that home ground advantage tends to have a greater effect in determining a winner than in increasing the difference between the goals scored by team i and j.

But, in contrast with the results of Pollard (1986), no differences between home ground advantages across English League divisions were found. This result is similar to that obtained by Dowie (1982). Therefore, arguing that the lower divisions have smaller crowds than the top ones and that there is no variation of the home ground advantage across the divisions, they concluded that the crowd's size has no effect on the home ground advantage.

Bland and Bland (1996) did not agree with this conclusion telling that playing before a crowd of 10'000 for a player in lower divisions may be as intimidating as playing before a crowd of 30'000 for a player in a major division. Therefore, they

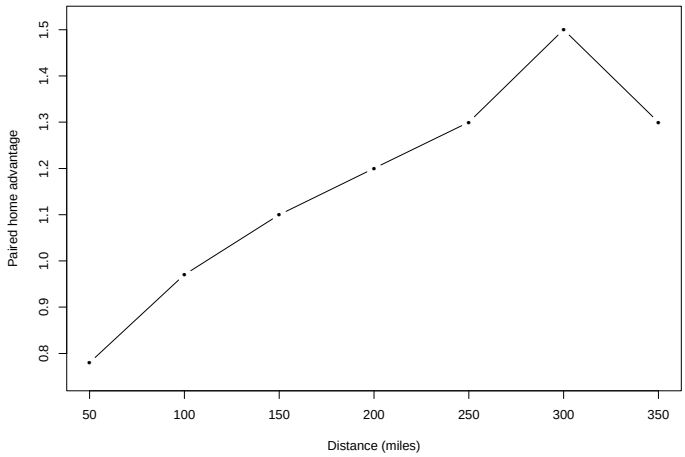

Figure 3: Paired home ground advantage compared to the distance (in miles) between the clubs.

do not consider it as a reasonable measure of the crowd effect on the home ground advantage.

Bland (1995) propose to consider the relationship between crowd size, points and goal difference for home and away matches. He argued that if attendance were more closely related to home points and goal difference than to away points or goal difference, this would show that attendance affects the team's performance. Calculating the correlation between home points and attendance, away points and attendance, home goal difference and attendance, and away goal difference and attendance, he concluded that the difference between home points and away points was positively related to average attendance. Therefore the crowd's size is a cause of home ground advantage.

Clark and Norman (1995) answered Bland and Bland (1996) by noting that if a team plays well, accumulating a lot of points or have a big goal margin, it will attract a bigger crowd than if it plays badly. Therefore it is possible that it is not the crowd which brings the good home team's performance at home but the result which brings the crowd.

In any case, by now, the relation between the crowd's size and the home ground advantage is not evident as a result of the difficulty to isolate this effect from the others; some further work is needed.

2.6 Artificial pitch surface

Another possible cause of home ground advantage is the fact to play on a surface other than lawn. During the years 1980–90, four English clubs installed an artificial pitch surface at their home ground. The English Football League asked Barnett and Hilditch (1993) to perform a statistical approach in order to find whether the artificial pitch gives an advantage for the teams playing at home or not.

They measured the performances of these four teams by means of points, goals and match results. Being aware that their measures do not hold any account of the team skills, they found that the teams with an artificial pitch surface had a little advantage. Hence artificial pitch surfaces cause larger home ground advantages. Therefore the Football English League forbade artificial turf after the publication of Barnett and Hilditch (1993).

2.7 Consequences of a red card

In soccer, contrary to hockey or hand ball, players could be punished for the rest of a match if they performed illegal actions like repeated flagrant fouls or preventing an adverse goal by illegal means: they receive a red card.

In order to modelise the effect of a red card on the outcome of a match, let us consider the two following assumptions:

- the two teams score according to two independent Poisson processes; and
- during the match there are different scoring intensities depending on the period of time. Indeed at the end of the match the team which is loosing will attack more intensively than at its beginning. This assumption is due to Morris (1981), who observed that the rate of scoring increases monotonically over the match.

According to these two assumptions, Ridder et al. (1994) found that a red card received early in the match increases considerably the 11-players team's chance to win, and decreases even more the 10-players team's chance to win; see Table 1. The last row of Table 1 also shows that if no team receives a red card during the match, the chances to win the match are equal for two teams of same skills. But if one of them receives a red card the probability for the team of 10 to win decreases. Note also that red cards may be an indication of the role played by chance in determining the outcome of a soccer match: receiving a red card depends on the referee's judgement of the foul committed and sometimes on what he did not see, because until now he has no right to watch the video of the match to take a decision.

2.8 Playing strategies

As seen in Section 2.3, chance is not the only winning factor: skill plays also an important role. Player's skills are necessary but good playing strategies may be

Minute of the red card	P(*team of 11 wins*)	P(*draw*)	P(*team of 10 wins*)
0	.65	.17	.18
15	.62	.18	.20
30	.58	.20	.22
45	.54	.21	.25
60	.49	.23	.28
75	.44	.24	.32
90	.375	.25	.375

Table 1: Probabilities of the outcome of the match by minute of receipt of the red card, for two teams of equal strength.

crucial to win as well.

The most important problem to evaluate different playing strategies is the lack of quantitative recorded data of all moves done during a soccer match. To gather such data, methods have been developed by Franks (1988) in Canada, Church and Hughes (1987) in England or Paukku (1994) in Finland. There had not been important statistical analysis of the recolted data since Ali (1988) and Pollard et al. (1988), but the work of Reep and Pollard (1997) became the reference in this field.

The basic measure to study different strategies is the team's ball possession: it starts when a team takes possession of the ball and ends when the ball runs out of play or when an opponent gains control of it, or when the referee stops the play because of a transgressed rule. The effectiveness of a team's ball possession depends on its outcome. Indeed a good possession will end up by a goal or at least a shot at goal, while a bad one will end by the loss of the ball. Furthermore, as all shots have not the same probability to score, it is a good idea to assign a weight to each shot, according to their probability to score for example. To obtain such weights, we refer the interested reader to Olsen (1988) and Pollard (1995) who studied this probability by means of several quantifiable factors.

To classify the different team's ball possessions, Reep and Pollard (1997) began to divide the field into six zones of play according to Figure 4. Then they differentiated between two types of possession: 'open play' and 'set play' (such as corners). Finally they define the yield y_{ij}, where i indicates the zone from where the action starts and j equals 1 for 'open play' and 2 for 'set play'. The yield is the probability of scoring a goal from zone i during possession type j minus the probability of receiving one. For example, y_{31} would be the yield of a possession originating in zone 3 as 'open play'. Using the 5'844 team possessions recorded during the 1986 FIFA World Cup in Mexico expected yield values were calculated; see Table 2. For example, considering 1'000 possessions originating in zone 4 as 'open play' a team could expect to score 10.9 more goals than it would receive. In all zones the expected yields for ball possessions originating as 'open play' are higher than for possessions originating as 'set play'. This is probably due to the extra time the defence has to position

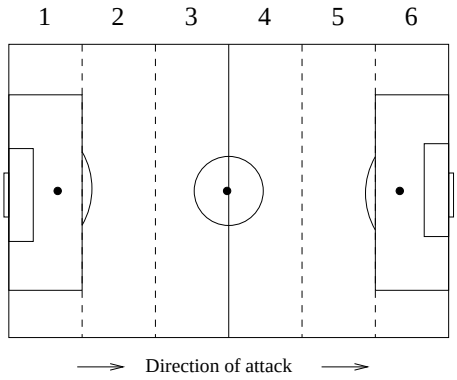

Figure 4: Division of the field of play in six zones.

itself when a 'set play' occurs. Furthermore, the yield obtained when the possession begins in zone 6 as 'open play' (78.3) illustrates that it is very important not to loose the ball in this area because otherwise it will result in one more goal received than scored, and this every 13 possessions. Considering that possessions originated as 'set play' in this area gave four times less goals. The strategy to adopt for the defence when endangered in zone 1 is certainly to shoot the ball out of the field limits in order to create a set play instead of loosing the ball in 'open play'.

It may be surprising that the yield depends so strongly on the origin of the possession as it is characterising the end of it. This dependence is due to the very short duration of the majority of the possessions. Indeed, analysing more than 23'000 possessions in the English first division, Reep et al. (1971) observed that fewer than 5% consisted of four or more completed passes. This may certainly explain this strong dependence.

Finally, instead of considering an average yield in each zone, Reep and Pollard (1997) considered the yield of specific actions, as in Table 3 for the 1986 FIFA World Cup. It is interesting to observe that long goal kicks have a negative yield meaning that such a strategy will procure more goals received than scored, therefore

Zone of origin	*Yield for 'open play'*	*Yield for 'set play'*
1	5.9	2.2
2	8.5	0.5
3	6.2	2.2
4	10.9	8.5
5	24.8	12.6
6	78.3	18.0

Table 2: Yield by 1'000 team possessions, classified by zone of origin and type of possession ('open play' or 'set play').

Situation	*Strategy*	*Yield*
Goal kick	Long	–2.7
Throw-in in own half	Short	–0.2
Possession in zone 4	Short passing only	11.1
	Running with the ball	16.3
	Long forward pass	23.1
Free kick in zone 5	Direct shot	12.5
	Other	16.8
Throw-in in zone 6	Short	3.5
	Long towards goal mouth	21.7
Centres from zone 6	Above waist height	33.3
	Below waist height	96.6

Table 3: Yield per 1'000 team possessions from playing strategies in different situations for the 1986 FIFA World Cup.

it is certainly better to avoid it. Table 3 may also lead us to conclude that in zone 4 long forward passes are a good choice, which seems to agree with the nature of English soccer: throw-ins in zone 6 have to be long and directed towards the goal mouth while centres coming from this zone are more effective below the waist than above it. Note that these results were obtained from the 1986 FIFA World Cup and that soccer may have evolved by now so that some results are not accurate any more. The same analysis may be repeated for actual competitions gathering enough data. To conclude their paper, Reep and Pollard (1997) noticed that *'soccer coaches, players, fans and the media are deeply sceptical and often suspicious, to the point of paranoia, at the suggestion that a statistician might have something useful to offer in the way of tactical analysis'*. Hence there is certainly a lot of work to be done in order to obtain more accurate results, but it will need a lot more time to convince the main actors that statistical and tactical analysis get along.

3 Round Robin tournaments

3.1 Introduction

Round Robin tournaments are competitions where every team plays every other team an equal number of times and after every match a different quantity of points is delivered whether the team wins, draws or looses. The team having the biggest total of points at the end of the tournament wins the competition. For example, all National championships are Round Robin tournaments.

This section discusses different methods to modelise soccer scores. Some resulting models predicting the outcome of Round Robin tournaments are presented as well.

3.2 Poisson and negative binomial distributions

Moroney (1956), followed fifteen years later by Reep et al. (1971), observed that the negative binomial distribution fits closely to the number of goals scored by a specific team during a match. In spite of this, Maher (1982) used a Poisson distribution to modelise the number of goals scored by teams i and j, in accordance with Ridder et al. (1994). Hence it may be quite important to know which of these two distributions fits soccer matches' scores best in order to develop new models for Round Robin tournaments using the adequate distribution.

Baxter and Stevenson (1988), from demand of the manager of an English soccer team, fitted the two distributions on scores between 1946 and 1984. Using the 5% and 1% critical values of the negative binomial and the Poisson distribution they found that, before 1970, the negative binomial distribution is more accurate than the Poisson distribution, but after 1970, either distribution is usually adequate. This separation into two periods coincides with the decline of goals observed in the 1960's. Subsequently, Baxter and Stevenson (1988) confirm the results found by Moroney (1956) and Reep et al. (1971), telling that before 1970 the negative binomial distribution fitted closer the data than the Poisson distribution. They also agreed with Maher (1982) and Ridder et al. (1994), observing that after 1970 the two distributions are adequate.

3.3 Soccer standings

Knowing which distribution to use, it is important to determine the parameters of this distribution and therefore why a team scores.

It certainly mainly depends on the team's ability but it may also depend on the team's will to stress its attack or defence. These changes of strategy are due to the urgency of scoring in order to win the match, as observed by Croucher (1984), or to improve its standing compared to a team which has equal points; a team other than the actual adversary. The criteria used to separate the two teams will be determinant for the choice of the strategy. By now it is the difference between the goals scored

and the goals received. Note that when the teams have the same goal difference the biggest number of goals scored becomes determinant. The question asked in public places, where soccer fans meet before and after the matches, is whether this criteria is the best or whether a ratio between the goals scored and the goals received would be better.

The difference between these two notions may not be evident at first sight. An easy way to illustrate it is to notice that a 3–1 victory is the same as a 6–4 victory using a 'difference' criteria, while it is the same as a 6–2 victory using a 'ratio' criteria. This difference may have a great influence on the teams' strategy. Indeed, using the ratio, a team loosing 3–1 will be encouraged to score because it may double its ratio with only one goal scored, while the opposing team would have to score three more goals to obtain the ratio it had before. Therefore, in such a case, the winning team will certainly be more defensive than it would be using the 'difference' criteria. Wright (1997) observed this fact by analysing the final table of the 1995–96 English Premier League season. He calculated the ratio between the goals scored and the goals received; see Table 4. This table has four blocs in which the teams have equal points. In the first block, Aston Villa and Arsenal are tied on points (63 each) and on goal difference (17). Then the higher position in the final table goes to the team that scores more goals, which is Aston Villa. But using a 'ratio' criteria, Arsenal would have been placed on top, because it received less goals, and

Team	*Wins*	*Losses*	*Draws*	*Points*	Goals			
					For	*Against*	*Difference*	*Ratio*
Manchester United	15	6	7	82	73	35	38	2.086
Newcastle	24	8	6	78	66	37	29	1.784
Liverpool	20	7	11	71	70	34	36	2.059
Aston Villa	18	11	9	**63**	52	35	**17**	1.486
Arsenal	17	9	12	**63**	49	32	**17**	1.531
Everton	17	11	10	*61*	64	44	20	1.455
Blackburn	18	13	7	*61*	61	47	14	1.298
Tottenham	16	9	13	*61*	50	38	12	1.316
Nottingham Forest	15	10	13	58	50	54	–4	0.926
West Ham	14	15	9	51	43	52	–9	0.827
Chelsea	12	12	14	50	46	44	2	1.045
Middlesbrough	11	17	10	**43**	35	50	–15	0.700
Leeds United	12	19	7	**43**	40	57	–17	0.702
Wimbledon	10	17	11	41	55	70	–15	0.786
Sheffield United	10	18	10	40	48	61	–13	0.787
Coventry	8	16	14	*38*	42	60	–18	0.700
Southampton	9	18	11	*38*	34	52	–18	0.654
Manchester City	9	18	11	*38*	33	58	–25	0.569
Queen's Park Rangers	9	23	6	33	38	57	–19	0.667
Bolton	8	25	5	29	39	71	–32	0.549

Table 4: English Premier League standings for the 1995–96 season.

therefore would have had a better seeding for the European competitions. In the last block, Coventry and Southampton are in the same situation using the 'ratio' criteria. Coventry would have been placed on the top because it had scored more goals. This confirms the general strategy pattern found above: with a 'ratio' criteria the best teams have to be defensive while the teams with a negative goal difference have to score in order to improve their ratio.

Considering the influence of both criteria on the goal scoring strategy, the goal distribution parameter will first depend on the team's attack and defence but also on the game's stake. Naturally the home ground advantage and the other factors observed in Section 2 have their importance too.

Now let us describe some methods used in order to calculate the parameters of the goal scoring distribution.

3.4 Maximum likelihood method

Assuming that the goal scores are Poisson distributed, Keller (1994) first proved that the probability for team i and team j to draw is

$$\mathrm{P}(i \text{ ties } j) = \frac{\mathrm{d}}{\mathrm{d}\lambda}\mathrm{P}(i \text{ beats } j),$$

where λ is the parameter of the goal scoring distribution of team i. This result does not hold if the scoring distribution is not Poisson. He wanted to calculate the probability for England to win, tie or loose when it faces Scotland. In order to estimate the distribution's parameter λ of England and the Scotland's one μ, he used the maximum likelihood method. He found the following estimators: $\hat{\lambda} = S_e/N$, $\hat{\mu} = S_s/N$, where N is the number of games played, and S_e and S_s the number of goals scored by England and Scotland respectively. Considering $N = 98$ games between the two teams where $S_e = 179$ and $S_s = 164$, he found that England will win 42% of the matches, draw 22% and loose 36%.

Unfortunately this method does not take into account the home ground advantage. It also demands a lot of data to obtain good estimators because it always considers only the matches played between the two teams involved. Therefore estimating the result of a Round Robin tournament knowing the first half will be difficult, because each team would have played only one time versus each other.

3.5 Generalised linear model

Using independent Poisson variates characterising team i and team j's goal scores, Kuonen (1996) and Lee (1997) used a generalised linear model in order to estimate λ and μ, respectively team i and j's goal scoring distribution parameters. It yields

the following model:

$$\begin{aligned} \log(\lambda) &= c + \gamma + a_i + d_j \text{ and} \\ \log(\mu) &= c + a_j + d_i, \end{aligned}$$

where c is a constant which expresses the average score in a match, γ is the home ground advantage value, and a_i and d_i are representing the attack and defence skills, $\forall\, i \neq j$. This model takes into account the home ground advantage, and gives also an attack and a defence skill to each team which permit to simulate different situations as, for example, the expulsion of a player for several matches by reducing the team's attack or defence parameters. Finally, it needs less data than the method proposed by Keller (1994), because every match may be considered in order to estimate all the parameters of each team. Kuonen (1996) used this model to simulate the Italian, French and German National championships from 1993 to 1995. He noticed that there were years where the model fitted better the data than during other years and that it was preferable to bet during the end of the championship using such a model. Lee (1997) simulated the 1995–96 English Premier League season and concluded that Manchester was lucky to win and that Liverpool certainly deserved to finish second, instead of Newcastle United.

3.6 Time dependent models

3.6.1 Improving Maher (1982)'s model

The next step to perform, in order to improve the prediction models, is to introduce a time dependence. Indeed every model observed by now had static parameters, i.e. teams were assumed to have a constant performance rate over time, whereas in fact team performances are varying through time.

Let us consider the model developed by Dixon and Coles (1997), who improved Maher (1982)'s model (see Section 2.4 for a complete description) in two ways.

- First, they observed that the assumption of independence between scores is no more reasonable for low scores as 0–0, 1–0 or 0–1. Therefore they introduced a function $\tau_{\lambda,\mu}(x, y)$ depending on x and y, the number of goals scored by home team i and team j. This function was added to the joint Poisson distribution as follows. Let $\lambda = \alpha_i \beta_j \gamma$ and $\mu = \alpha_j \beta_i$ be the distribution parameters of team i and j respectively, then

  $$\mathrm{P}(X_{i,j} = x, Y_{i,j} = y) = \tau_{\lambda,\mu}(x, y)\mathrm{P}(X_{i,j} = x)\mathrm{P}(Y_{i,j} = y)$$

  and

  $$\tau_{\lambda,\mu}(x, y) = \begin{cases} 1 - \lambda\mu\rho, & \text{if } x = y = 0, \\ 1 + \lambda\rho, & \text{if } x = 0, y = 1, \\ 1 + \mu\rho, & \text{if } x = 1, y = 0, \\ 1 - \rho, & \text{if } x = y = 1, \\ 1, & \text{otherwise.} \end{cases}$$

In this model ρ is a dependence parameter included between $\max(-1/\lambda, -1/\mu)$ and $\min(1/\lambda\mu, 1)$.

- The second change concerned the time dependence of the model. Let t_k be the time when match k was played. Dixon and Coles (1997) introduced a function $\phi(t-t_k)$ depending on the difference between now (time t) and when the match k was played (time t_k). The $\phi(t-t_k)$ is a weight for the joint distribution as follows:

$$\begin{aligned} \mathrm{P}(X_{i,j,k} = x_k, Y_{i,j,k} = y_k) &= \{\tau_{\lambda,\mu}(x_k, y_k)\mathrm{P}(X_{i,j,k} = x_k) \\ &\quad \times \mathrm{P}(Y_{i,j,k} = y_k)\}^{\phi(t-t_k)}. \end{aligned}$$

  The latter may be interpreted as: the more ancient the match k was, the less important it is to estimate the outcome of the next match. Dixon and Coles (1997) used the function $\phi(t) = \exp(-\xi t)$ with $\xi > 0$.

Their model was created in order to enable to bet with a positive expected outcome. Performing several statistical tests, they demonstrate that this goal has been achieved.

3.6.2 Using paired comparisons

Another time dependent model is the one created by Fahrmeir and Tutz (1994) using an extension of the method of paired comparisons. First developed by Bradley and Terry (1952), this method can be used to consider every team in a Round Robin tournament and to compare each team with each other in pairs. Let $y_{i,j}$ denote the result of a match between the pair (i, j) of teams. And $y_{i,j} = 1$ if team i wins and $y_{i,j} = 2$ if team j wins. Then the model is

$$\mathrm{P}(y_{i,j} = 1) = \frac{\exp(\alpha_i - \alpha_j)}{1 + \exp(\alpha_i - \alpha_j)},$$

where α_i represents the strength of team i, $\forall\, i$. Fahrmeir and Tutz (1994) made two extensions to adapt it to soccer.

- First, they let $y_{i,j}$ take values from $\{1, \ldots, k\}$ with $k \geq 2$. This permits to create a grading dominance of the score between the two teams: 1 means that i dominates j the most strongly and k means the opposite. For example, if $k = 3$ than the categories 1, 2 and 3 mean i wins, draws and j wins respectively.

- Second, in Bradley and Terry (1952)'s model, the pairs (i, j) and (j, i) were the same. Regarding the home ground advantage this cannot be in a soccer model. Therefore taking into account the specific order in the pairs would be necessary.

Fahrmeir and Tutz (1994)'s basic assumption is that for each team i there exists a random variate $U_i = \alpha_i + \epsilon_i$ where ϵ_i is a random variable. The connection between U_i and $y_{i,j}$ is

$$y_{i,j} = r \Leftrightarrow \theta_{r-1} < U_j - U_i < \theta_r,$$

where $\theta_0 < \cdots < \theta_k$ are thresholds. Assuming a continuous distribution $\mathrm{F}(\cdot)$ for the differences $\epsilon_j - \epsilon_i$ they obtain the general ordinal paired comparison model

$$\mathrm{P}(y_{i,j} = r) = \mathrm{F}(\theta_r - \alpha_i - \alpha_j) - \mathrm{F}(\theta_{r-1} - \alpha_i - \alpha_j).$$

The next step was the introduction of time dependence into the model. The first time dependent model they considered was simply a time dependent version of the previous one, i.e. at time t

$$\mathrm{P}(y_{i,j}^{(t)} = r) = \mathrm{F}(\theta_{t,r} - \alpha_{t,i} - \alpha_{t,j}) - \mathrm{F}(\theta_{t,r-1} - \alpha_{t,i} - \alpha_{t,j}),$$

including relations to obtain $\alpha_{t+1,i}$ from $\alpha_{t,i}, \forall\, t, i$. They continued to improve the time dependent model using a generalised Kalman filter for paired comparisons, for which we encourage the interested reader to have a look at Fahrmeir and Tutz (1994) in order to obtain more information. Finally, they applied their model to the German National soccer league from 1965 to 1987. What is interesting in their results is that the fluctuations of Bayern Munich's ability through years is in good agreement with reality. Indeed, they obtained a peak in 1970–72 corresponding to the team's most successful years when Franz Beckenbauer was captain. Then the ability curve began to decrease illustrating the departure of Franz Beckenbauer and the other principal players. And then the curve increased up to the late 1980s, characterising the time it needed to reform a strong team. This model shows a remarkable fit to the variating ability of each team throughout time, whereas it assumes good statistical knowledge and the need of powerful computers.

3.6.3 Using Markov chain Monte-Carlo methods

The last model we consider in this section is the model developed by Rue and Salvesen (1999) using Markov chain Monte-Carlo (MCMC) methods. Their model is based on Lee (1997); see Section 3.5. They modified the model as follows.

- Instead of defining a constant γ measuring the home ground advantage, they replaced Lee (1997)'s constant c by $c^{(x)}$ and $c^{(y)}$, where x and y are the number of goals scored by the home and away teams. Hence $c^{(x)}$ is a constant describing the average number of goals scored at home while $c^{(y)}$ describes the number of away goals.

- They included a psychological effect: if team i is stronger than team j than team i will tend to underestimate team j. Let $\Delta_{i,j} = \{a_i + d_i - (a_j + d_j)\}/2$

measure the difference in strengths between team i and j. Then

$$\begin{aligned} \log(\lambda) &= c^{(x)} + a_i + d_j - \kappa\Delta_{i,j} \text{ and} \\ \log(\mu) &= c^{(y)} + a_j + d_i + \kappa\Delta_{i,j}, \end{aligned}$$

where $\kappa > 0$ is a small constant indicating the intensity of the psychological effect.

As Dixon and Coles (1997) (see Section 3.6) they noticed that for low scores, the assumption of independent Poisson variates was not reasonable. Hence they used the same correction factor $\tau_{\lambda,\mu}(x,y)$ with a dependence parameter $\rho = 0.1$. They obtained the same distribution as Dixon and Coles (1997), which they improved in two ways.

- When one of the teams scores many goals, it is certainly highly demotivating for the other team and it is in contradiction with the assumption that the goal scoring intensity does not depend on the goals scored in the match. Therefore they truncated the Poisson distribution after 5 goals. So the score of 7–0 or 6–5 will be interpreted as 5–0 or 5–5. The goals scored after the fifth are uninformative for the two team's properties. Let us denote by π^* the resulting truncated law.

- They suggested that the match result is less informative than suggested by π^*. Therefore they constructed a more robust model by forming

  $$\pi(x,y|\lambda,\mu) = (1-\epsilon)\pi^*(x,y|\lambda,\mu) + \epsilon\pi^*(x,y|\ \exp(c^{(x)}),\ \exp(c^{(y)})),$$

  where the $(1-\epsilon)$ part determines the important information while the ϵ part follows the law of an average match, $0 \leq \epsilon \leq 1$

The last step is to introduce the time in the model. They chose to use a Brownian motion to tie together a_i at the two time points t' and t''. Inference was performed using a MCMC algorithm. For more precisions, we invite the interested reader to have a look at Rue and Salvesen (1999). Finally they applied their model to the second half of the 1997–98 English Premier League season in order to develop a betting scheme and answer questions about the final ranking table of the tournament. Their results include: Manchester United, and not Arsenal, would have deserved to win the championship, or Aston Villa could easily have been 15th instead of 7th. They also obtain interesting results concerning the time dependence: the decreasing attack skill of Manchester United during the second half of the championship, certainly due to the injury of Denis Irwin.

To conclude we observe that such time dependent models permit to obtain better results than the previous ones but are certainly complicated to use. Note that the principal advantage of the MCMC model is that it does not only give the outcome

of the match but also the final score of it, which is important to classify teams with the same number of points as seen in Section 3.3.

This section showed us that there is a great passion to develop more and more complicated models for Round Robin tournaments whereas we will see in Section 4 that few statisticians developed models for Knock Out tournaments.

4 Knock Out tournaments

4.1 Introduction

The Knock Out tournaments are competitions where every match can have only two outcomes: win or loose. All the teams are first classified and than each team plays the team which is beside it. In the National cups and the second part of the FIFA World Cup, the team which loose the game is definitely eliminated from the tournament while in the European leagues, as the UEFA Cup, the former Cup Winners Cup or the second part of the Champions League, each team plays two games versus its opponent, one at home and the other away. The best team of the two matches follow the competition.

This section will present a model developed by Kuonen (1997) in order to determine the possible winners of a Knock Out tournament. Only his model will be presented because we did not find others on this subject.

4.2 Seeding coefficients

The most important thing in a Knock Out tournament is the seeding. With a good seeding you may play several rounds before confronting one of the best teams, therefore your progression would be easier. It seems natural that the seeding coefficients depend on the strength of the teams in order not to oppose the two best teams during the first round.

Kuonen (1997) compared three methods permitting to calculate seeding coefficients.

- The first one assumes that these coefficients (which represents the team strengths) stayed constant during the tournament. Their value is determined by results of the last three years in a Knock Out tournament, allowing two points for a win, one for a draw and an additional point for reaching the quarter final, the semi final or the final. Then the coefficient is calculated by means of

$$\text{coefficient} = \frac{\text{points achieved during the three year period}}{\text{games played during the three year period}}.$$

- The second method assumes that the seeding coefficients change at each round. The initial coefficients are calculated following the previous method and for the following, he used the relation below. Let $C_{old,i}$ be the old coefficient of team i and $C_{new,i}$ its new one, then

$$C_{new,i} = C_{old,i}\left(1 + \frac{C_{old,O(i)} + \gamma}{C_{old,i} + \gamma}\right),$$

where $\gamma \geq 0$ is a constant and $O(i)$ is the most probable opponent of i for the new round, i.e. the team among the possible opponents of i which had the higher probability to win the previous round.

- Instead of considering only one possible opponent, the third method considers every possible opponent of team i in oder to calculate its new seeding coefficient. Therefore, letting $r_j = (C_{old,j} + \gamma)/(C_{old,i} + \gamma)$, the new coefficient became

$$C_{new,i} = C_{old,i} \sum_{j \in \mathcal{O}} (1 + r_j) \mathrm{P}(j \text{ wins in the previous round}),$$

where $\mathcal{O}$ is the set of all possible opponents. To use the two last methods, the knowledge of the probability of winning in a specific round is needed. This probability will be discussed in the next section.

4.3 Logistic regression model

Kuonen (1997) used a logistic regression model in order to calculate the probability $\mathrm{P}_k(i, j)$ of team i to win the kth round against team j. This model is closely related to the Bradley and Terry (1952) model seen in Section 3.6.2. It gives the following probability

$$\mathrm{P}_k(i, j) = \frac{\exp\{\alpha + \beta(C_{k,i} - C_{k,j})\}}{1 + \exp\{\alpha + \beta(C_{k,i} - C_{k,j})\}},$$

where α and β are two regression constants, and $C_{k,i}$ is the seeding coefficient of team i during the kth round. Therefore, the probability that team i wins in round k is

$$\begin{aligned} \mathrm{P}(i \text{ wins in round } k) &= \mathrm{P}(i \text{ wins in round } k-1) \\ &\quad \times \sum_{j \in \mathcal{O}} \mathrm{P}_k(i, j) \mathrm{P}(j \text{ wins in round } k-1). \end{aligned}$$

In order to determine which of the three methods is the best, Kuonen (1997) applied the model to European soccer cup tournament data from 1992 to 1996, calculating the probability of each team to win the competition. Considering 376 matches, the best method found was the one which assumes the constancy of the teams coefficients, therefore of the team strengths. This method was better because it was faster and easier to apply than the two others, and because the other methods did not give better results. This result is surprising because the ability of each team is subject to a lot of variation in reality, caused by the transfers, injuries or suspensions. Kuonen (1997) predicted in average 64.49% of the games correctly. It is a good result but more work may be done in this field in order to find a model which predict correctly more than 64.49% of the game using variation of team's strength throughout the rounds.

To conclude, let us notice that using a simplified version of this model applied to the second half of the 1998 France World Cup, Kuonen et al. (1999) predicted that France would win the final with a probability of 57.4%, whereas Brazil was 3–1 favourite at the bookmakers odds.

5 Knock Out tournaments with preliminary Round Robin stage

5.1 Introduction

This section speaks about tournaments like the FIFA World Cup or the Champions League which have both a Round Robin and a Knock Out part. Therefore the models seen in the last two sections may be used to determine the winning team. We will principally talk about the influence of the seedings in the different groups during the first half and the second half of the FIFA World Cup. The Champions League will not be studied because of the lack of articles about it.

5.2 Is there a group better than the others?

The FIFA World Cup is divided into two parts.

- In the first part the teams are gathered in six groups: A to F. In each of the group a Round Robin tournament is performed. Two of the first seeds are reserved to the actual World Champion and the host team. The other teams are classified following their rank calculated by the means of the FIFA-Coca Cola world soccer ratings system; see Stefani (1997) for a complete description of this rating system.
- The second part is a Knock Out tournament where the seeding depends on the results of the Round Robin part as explain in Table 5. For example, A–1 represents the first ranked team in group A after the Round Robin part. Only the four best third ranked teams play the Knock Out part of the tournament.

McGarry and Schutz (1994) studied the behaviour of the FIFA World Cup structure in order to know whether beginning the competition in one group is more beneficial than beginning in an other. To investigate on this, they gave a rating score to each team involved in order to modelise their strengths. The first seeds inherited a score of 100, the second seeds a score of 80, the third a score of 60 and the final seed a score of 40. These scores do not vary throughout time. A win is awarded between

Directly seeded	B–1	C–1	A–1	D–1	A–2	F–1	F–2	E–1
Possible opponents	A–3 C–3 D–3	A–3 B–3 F–3	C–3 D–3 E–3	B–3 E–3 F–3	C–2	E–2	B–2	D–2

Table 5: Seeding of the Knock Out part considering the results of the Round Robin part of the FIFA World Cup.

team i and team j, following a paired comparisons model (see Bradley and Terry (1952)) as follows

$$\mathrm{P}(i \text{ beats } j) = \frac{R_i}{R_i + R_j},$$

where R_i is the rating score of team $i, \forall\, i \neq j$. For the simulation a Monte Carlo procedure was used. For their analysis, they first checked whether the tournament is balanced or not by giving each team the same rating score. The results revealed that it is not, because of the choice of the four third ranked teams. Therefore they resolved this problem in the simulation by introducing a random order for pool access from which the best third ranked teams are elected. Using this model they simulated the competition using the different rating scores seen above. They found an inequitable structure, principally due to the seeding of the Knock Out part. Indeed, the A, B, C and D groups are directly favoured because their first ranks have to play third ranked teams, whereas the first ranks of group E and F are matched to second ranked teams. It is further noticeable that the second ranked team of group E has a particularly difficult match versus a first ranked team. Finally, putting the rating score of a first seed to 120, they tried to determine whether there is a group better then the others for a first seed. The results were that the groups A and C were favourable for the rest of the competition whereas the groups E and F were not.

5.3 Is the host team's first seed an advantage?

Another interesting behaviour that McGarry and Schutz (1994) analysed was the importance of the host team's first seed. They imagined a weak host team H with a rating score of 40. Let us introduce a new notation: (1)D represent the first seed of group D. During the simulation, H was first seeded (4)D and obtained a score of 1275 points. Then it was seeded (1)A and its score increased to 1473 points. The score obtained the same increase whether H was seeded (1)A, (1)B, (1)C or (1)F, but did not change for the seeds (1)D and (1)E. When H was seeded (1)D from (4)D the score did not change because H and the previous (1)D just changed their places in group D, therefore there were always teams with ratings of 100, 80, 60, 40. But when H became the first seed of another group than D, this had for effect to strengthen group D by adding another team with a rating score of 100. Hence when H became (1)E the first seed advantage is negated because, during the first round of the Knock Out part, H will meet D–2 (E–1 meets D–2) or F–1 (E–2 meets F–1) which are two teams with a high probability to have a rating score of 100. Therefore the advantage of the promotion of H to a first seed may be negated by the changes occurring in the seedings of the other groups.

5.4 Was France's World Cup win pure chance?

Finally, let us describe a model which is able to predict the winner of the FIFA World Cup. Kuonen and Roehrl (2000) created a simple model divided in two parts, one model for each part of the competition. For the Round Robin tournament, they estimated the outcome of each match using Stefani (1980)'s model (described in Section 2.5) without considering any home ground advantage, because except the host team every other team plays on neutral ground. Then the outcome of a game between i and j is given by

$$w_{ij} = u_i - u_j + \epsilon_{ij}, \forall\, i \neq j.$$

They used least squares estimators in order to find the teams ratings. For the second part of the competition, the Knock Out part, they used the model described in Section 4 which gave the probability for team i to win in round k by

$$\begin{aligned} \mathrm{P}(i \text{ wins in round } k) &= \mathrm{P}(i \text{ wins in round } k-1) \\ &\quad \times \sum_{j \in \mathcal{O}} \mathrm{P}_k(i,j)\mathrm{P}(j \text{ wins in round } k-1). \end{aligned}$$

They compared the least squares ratings to other ratings like the FIFA-Coca Cola world soccer ratings. The FIFA ratings gave a probability over than 50% for a victory of Brazil over all other teams, while the least squares ratings' best winning probability was for France, which is in accordance with Kuonen et al. (1999). Therefore, considering that the least squares ratings were obtained only using the information from the Round Robin part, it is certainly better to consider just the most recent data in order to have an accurate model. The data from earlier years seem to have a bad influence on the other ratings estimations.

6 Conclusion

Although initially the statistical community did not like the mix of statistics and soccer, this report have shown that, since Moroney (1956), the statisticians became enthusiastic concerning soccer. The fact is underlined through the publication of a lot of articles overviewing many soccer fields. The statistical methods employed to predict the outcome of soccer matches go from simple models as the maximum likelihood method to more complicated ones as MCMC algorithms or Kalman filters. These methods may become a good pedagogical support in order to teach statistics through examples which may interest everybody. Rather than creating predicting models, several statisticians interested themselves to other faces of soccer, like the advantage a team may obtain playing at home or the disadvantage to receive a red card or still the research of the best playing strategies. We also observed that concerning the Knock Out tournaments, few articles have been written. Therefore it is certainly a good starting point for anybody who wants to develop statistical models concerning soccer.

Finally, as the present report only talked about statistics in soccer, we invite the reader interested in statistical models concerning other sports to have a look at the excellent book edited by Bennett (1998).

Appendix: Web resources

This section will give to the interested reader some web links to soccer pages. These links are divided into two parts: the first part concerns soccer data and general information about soccer and the second part concerns statistics in soccer and betting.

First part

- The FIFA official web site.

  `http://www.fifa.com/index.html`

- The UEFA official web site.

  `http://www.uefa.org/`

- The results of all European cups.

  `http://www.z-axis.com/uefa/home.html`

- The results of the National championship of several European countries.

  `http://users.skynet.be/Ogfoot/index.html`

- Other data about soccer.

  `http://football.sports.com/`

Second part

- A list of all statistical journals .

  `http://www.maths.uq.oz.au/~gks/webguide/journals.html`

- Softwares on line in order to perform predictions.

  `http://users.aol.com/soccerslot/forthdim.html`

- A site permitting to bet on line.

  `http://www.betandwin.com/Bet/en/index.htm`

References

Ali, A. H. (1988). A statistical analysis of tactical movement patterns in soccer. In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), pp. 302–308. Spon, London.

Barnett, V. and Hilditch, S. (1993). The effect of an artificial pitch surface on home team performances in football (soccer). *Journal of the Royal Statistical Society A*, **156**, 39–50.

> **Abstract:** In the first four divisions of the English Football League four teams had their home matches on artificial pitch surfaces at a certain time over the last years. A statistical analysis of end-of-season results for the four divisions showed that there is an advantage gained by the home team on such pitches. This advantage is important enough to be a cause of concern.

Baxter, M. and Stevenson, R. (1988). Discriminating between the Poisson and negative binomial distributions: An application to goal scoring in association football. *Journal of Applied Statistics*, **15**, 347–354.

Bennett, J. (ed.) (1998). *Statistics in Sport.* Arnold Applications of Statistics, Arnold, London.

Bland, N. D. (1995). A mathematical analysis of football. *Mathematics Project.* Pimlico School, London.

Bland, N. D. and Bland, J. M. (1996). Comment on 'Home ground advantage of individual clubs in English soccer' (44, pp. 509–521). *The Statistician*, **45**, 381–383.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs 1: the method of pair comparisons. *Biometrika*, **39**, 324–345.

Church, S. and Hughes, M. (1987). A computerised approach to soccer notation analysis. *Abstract of the 1st World Congress of Science and Football, Liverpool*, p. 20. Liverpool Polytechnic, Liverpool.

Clark, S. R. and Norman, J. M. (1995). Home ground advantage of individual clubs in English soccer. *The Statistician*, **44**, 509–521.

Courneya, K. S. and Carron, A. V. (1992). The home advantage in sport competitions: a literature review. *Journal of Sport and Exercise Psychology*, **14**, 13–27.

Croucher, J. S. (1984). The effect of changing competition points in the English football league. *Teaching Statistics*, **6**, 39–42.

Dixon, M. J. and Coles, S. G. (1997). Modelling association football scores and inefficiences in the football betting market. *Applied Statistics*, **46**, 265–280.

> **Abstract:** A parametric model is developed and fitted to English league and FA Cup football data from 1992 to 1995. The model is motivated by an aim to exploit potential inefficiences in the association football betting market, and this is examined using bookmakers odds from 1995 to 1996. The technique is based on a Poisson regression model but is complicated by the data structure and the dynamic nature of team performances. Maximum likelihood estimates are shown to be computationally obtainable, and the model is shown to have a positive return when used as bases of a betting strategy.

Dobson, S. M. and Goddard, J. A. (1995). The demand for professional league football in England and Wales, 1925–92. *The Statistician*, **44**, 259–277.

Dowie, J. (1982). Why Spain should win the World Cup. *New Scientist*, **94**, 693–695.

Fahrmeir, L. and Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association*, **89**, 1438–1449.

> **Abstract:** When paired comparisons are made sequentially over time, it is natural to assume that the underlying abilities do change with time. Previous approaches are based on fixed updating schemes where the increments and decrements are fixed functions of the underlying abilities. The parameters that determine the functions have to be specified a priori and are based on rational reasoning. The authors suggest an alternative scheme for keeping track with the underlying abilites. Their approach is based on two components: a response model that specifies the connections between the observations and the underlying abilites, and a transition model that specified the variation of abilites over time. These two components form a non-Gaussian state-space model. Based on recent results, recursive posterior mode estimation algorithms are given and the relation to previous approaches is worked out. The performance of the method is illustrated by simulation results and an application to soccer data of the German Bundesliga.

Franks, I. M. (1988). Analysis of association football. *Soccer Journal*, **33**, 35–43.

Hill, I. D. (1974). Association football and statistical inference. *Applied Statistics*, **23**, 203–208.

> **Abstract:** A comparison of the final league tables of 1971–72 English soccer season with forecasts made by goals scored before the season began shows significant positive correlation. This seems to indicate that soccer results are not pure chance. The paper questions how the data would have been handled by statisticians who do not approve the results of significance tests.

Keller, J. B. (1994). A characterization of the Poisson distribution and the probability of winning a game. *The American Statistician*, **48**, 294–298.

> **Abstract:** The probability that a team with a certain mean score beats a team with a different mean score is calculated in the case where the score of each team is Poisson distributed; illustrating some conditions to the probability of a tie. The Poisson distribution is shown to fit some soccer data very well.

Kuk, A. Y. C. (1995). Modelling paired comparison data with large numbers of draws and large variability of draw percentages among players. *The Statistician*, **44**, 523–528.

Kuonen, D. (1996). Modelling the success of soccer teams in European championships. *Technical Report 96.1*, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Kuonen, D. (1997). Statistical models for knock-out soccer tournaments. *Technical Report 97.3*, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Kuonen, D., Chavez-Demoulin, V., Roehrl, A. S. A. and Chavez, E. (1999). La France, championne du monde de football: qui l'eût cru? *Technical Report 99.1*, Department of Mathematics, Swiss Federal Institute of Technology, Lausanne.

Kuonen, D. and Roehrl, A. S. A. (2000). Was France's World Cup win pure chance? *Student*, to appear.

Lee, A. J. (1997). Modeling scores in the Premier League: is Manchester United really the best. *Chance*, **10**, 15–19.

> **Abstract:** Assuming the distribution of the number of goals scored to be Poisson, the mean of the goals scored by the home team and the visiting team are estimated by means of a Poisson regression model; including home team advantage. The model was tested on the 1995–96 English Premier League season.

Maher, M. J. (1982). Modelling association football scores. *Statistica Neerlandica*, **36**, 109–118.

> **Abstract:** The author derives a model for soccer scores in a game between specific teams, accounting for the different quality of the teams involved. He obtained maximum likelihood estimates for a model in which the scores of the home and away teams, in any games, are independant Poisson distributions; means being functions of the respective team previous' performances.

McGarry, T. and Schutz, R. W. (1994). Analysis of the 1986 and 1994 World Cup Soccer Tournament. *ASA Proceedings from the 1994 Joint Statistical Meeting in Toronto, Statistics in Sports*, 61–65.

Abstract: The paper tries to answer different questions about the FIFA World Cup tournaments using Monte-Carlo simulation procedures. Is the home team's first seed ranking an advantage? Is there a group in the Round Robin part of the tournament in which the probability to win the World Cup is the best? In addition to answer these questions the authors also give a comparison between the different seedings in the Knock-Out part of the 1986 and 1994 Word Cup.

Moroney, M. J. (1956). *Facts from figures.* 3rd edition, Penguin, London.

Abstract: Study of the distribution of the goals scored in a soccer match. The conclusions are that the Poisson distribution provides an adequate fit to scores, but improvements could be obtained using the negative binomial distribution.

Morris, D. (1981). *The Soccer Tribe.* Jonathan Cabe, London.

Olsen, E. (1988). An analysis of goalscoring strategies in the World Championship in Mexico, 1986. In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), pp. 373–376. Spon, London.

Paukku, T. (1994). And it's another goal. *New Scientist*, 30–32.

Pollard, R. (1986). Home advantage in soccer: a retrospective analysis. *Journal of Sports Sciences*, **4**, 237–248.

Pollard, R. (1995). Do long shots pay off? *Soccer Journal*, **40**, 41–43.

Pollard, R., Reep, C. and Hartley, S. (1988). The quantitative comparison of playing styles in soccer. In *Science and Football* (eds. Reilly, T., Lees, A., Davids, K. and Murphy, W. J.), pp. 259–277. Spon, London.

Reep, C. and Benjamin, B. (1968). Skill and chance in association football. *Journal of the Royal Statistical Society A*, **131**, 581–585.

Abstract: Study of the probabilities to do a certain number of passes in a row in a phase of action during a soccer match. These probabilities were tested versus negative binomial distribution and it results a remarkable fit when all season matches were taken in account.

Reep, C. and Pollard, R. (1997). Measuring the effectiveness of playing strategies at soccer. *The Statistician*, **46**, 541–550.

Abstract: Using a notational system, which records on-the-ball events taking place throughout a soccer match, the game can be broken down into a series of team possessions. To assess the effectiveness of a team possession, a quantitative variable is developed representing the probability of a goal being scored minus the probaility of one being conceeded. This variable, called the yield, can be

used to evaluate both the expected outcome of a team possession originating in a given situation, as well as the actual outcome of the possession. In this way, the effectiveness of different strategies occuring during the possession can be quantified and compared.

Reep, C., Pollard, R. and Benjamin, B. (1971). Skill and chance in ball games. *Journal of the Royal Statistical Society A*, **134**, 623–629.

Abstract: The authors have followed up earlier work on soccer indicating that the negative binomial distribution would be applicable to certain movements or performances in other ball games by testing applications to cricket, ice hockey, baseball and lawn tennis. In Poissonian situations good fits were obtained. Poor fits were obtained in situations where individual skill appeared to play a stronger role.

Ridder, G., Cramer, J. S. and Hopstaken, P. (1994). Down to ten: estimating the effect of a red card in soccer. *Journal of the American Statistical Association*, **89**, 1124–1127.

Abstract: The authors investigate the effect of the expulsion of a player on the outcome of a soccer game by means of probability model for the score. They propose estimators of the expulsion effect that are independent of the relative strength of the teams. They use the estimates to illustrate the expulsion effect on the outcome of a match.

Rue, H. and Salvesen, O. (1999). Predicting and retrospective analysis of soccer matches in a league. *Technical Report*, Department of Mathematical Sciences, Norvegian University of Science and Technology, Trondheim.

Stefani, R. T. (1980). Improved least squares football, basketball and soccer predictions. *IEEE Transactions on Systems, Man and Cybernetics*, **10**, 116–123.

Stefani, R. T. (1997). Survey of the major world sports rating systems. *Journal of Applied Statistics*, **24**, 635–646.

Wright, D. B. (1997). Football standings and measurement levels. *The Statistician*, **46**, 105–110.

Abstract: This is a discussion about the way in which soccer standings are calculated in the English Premier League. The author compares two methods to classify the teams with the same points. First, he calcuates the interval between the goals for and the goals against, and, second, he calculates the ratio between the goals for and the goals against. Using the results of the English Premier League he shows that the two methods deliver different standings.